Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Semantic Search in an Academic Repository Using Sentence-BERT and FAISS

Ihsan Lubis [1)*], Husni Lubis [2)], Inaya Nur Wahidah [3)]
[1,2,3)]Information System, Faculty of Engineering and Computer Science, Universitas Harapan Medan, Indonesia
[1)] ihsan.lubis@gmail.com, [2)]husni.lubis82@gmail.com, [3)] inayawahidah@gmail.com

Submitted :Feb 26, 2026 | **Accepted** : Mar 26, 2026 | **Published** : April 2, 2026

Abstract: Academic repositories serve as centralized platforms for storing and managing scientific documents, including research papers, reports, and administrative records. Yet, traditional keyword-based search systems often struggle to deliver relevant results. These systems typically fail to capture the contextual meaning of user queries, which leads to mismatches when the query terms differ from those found in the documents. To overcome this limitation, this study introduces a semantic search approach for academic repositories by combining Sentence-BERT as the text embedding model with FAISS as the vector-based similarity search engine. In the proposed system, documents stored in a MySQL database are first preprocessed to remove HTML tags, then converted into semantic vector representations using Sentence-BERT. These vectors are indexed with FAISS, enabling fast and efficient similarity searches compared to conventional keyword matching. The system architecture integrates FastAPI as the backend service for indexing, searching, and evaluation, while CodeIgniter 4 functions as the frontend framework for document management by administrators and end users. Evaluation was carried out using three test sets, each containing ten queries. Performance was measured using Recall@K, normalized Discounted Cumulative Gain (nDCG), Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and search latency. Experimental results show that the system achieved an average Recall@K of 0.61, a MAP of 0.39, and a No-Hit rate of 0.033, meaning all queries successfully retrieved results. Although the nDCG value declined in the third test set, the system consistently returned relevant documents.

Keywords: Academic Repository; Semantic Search; Sentence-BERT; FAISS; Information Retrieval

INTRODUCTION

The rapid development of information technology has transformed how academic communities store, manage, and access scientific knowledge (Diana & Ekasari, 2021; Heriani et al., 2025). Academic repositories have become a key solution for organizing essential materials such as journal articles, research reports, undergraduate theses, master's and doctoral dissertations, as well as administrative records (Kadang & Nasaruddin, 2025). Beyond supporting data management, these repositories play a vital role in ensuring open and sustainable access to knowledge (Safira, 2021; Tupan & Rahayu, 2022). However, as the volume of stored documents continues to expand, a pressing challenge emerges: maintaining retrieval processes that are not only fast and accurate but also responsive to the specific needs of users. Prior studies indicate that keyword-based retrieval models, which rely mainly on lexical matching, may not fully represent the contextual relationships embedded in natural-language queries (Amur et al., 2023; Khan et al., 2022; Xiong et al., 2024). Variations in terminology, the use of synonyms, and differences in language frequently lead to search results that are less relevant. In academic settings, this issue becomes particularly critical, as it can prevent researchers, lecturers, and students from finding references that truly align with their topics of study. To address this challenge, a new approach is needed one that can capture contextual meaning in text, allowing search systems to recognize not only lexical similarity but also semantic similarity.

Recent advances in Natural Language Processing (NLP) have established Bidirectional Encoder Representations from Transformers (BERT) as a major breakthrough in semantic text understanding (Xu et al., 2022). BERT and its extensions, such as Sentence-BERT, are designed to generate vector representations that capture the semantic relationships among words within sentences and documents. Unlike traditional methods that rely primarily on term frequency, these models interpret context bidirectionally and model the relationships between tokens. As a result, Sentence-BERT is particularly effective in producing stable sentence embeddings,

*name of corresponding author

which allow for more accurate measurement of semantic similarity between texts. Current research trends highlight the growing adoption of this model across diverse applications including semantic search (Santander-Cruz et al., 2022), question answering (Acharya et al., 2022), and document clustering (Zhu et al., 2023) due to its efficiency and ability to capture deep contextual meaning. Managing large scale vector data requires technologies that can perform search operations both efficiently and quickly. FAISS (Facebook AI Similarity Search) has become a widely adopted solution in research and real world applications for tackling similarity search challenges in high dimensional data (Krisnawati et al., 2024). It offers a range of vector search algorithms, from exact methods to highly efficient approximate approaches designed for large datasets. Within academic repositories, FAISS enables efficient similarity search over large document collections and is widely adopted in scalable retrieval systems. Recent studies highlight that integrating FAISS with transformer based embedding models significantly improves the scalability and speed of search systems without compromising result quality (Wang, Zeng, et al., 2024; Zoupanos et al., 2022). Consequently, FAISS stands as a critical component in ensuring the practical implementation of semantic search systems in dynamic and continuously expanding repository environments.

Recent studies in information retrieval and natural language processing show that semantic search methods built on pre trained language models greatly enhance search relevance (Ghali et al., 2025; Karri & Jangam, 2024; Naqvi et al., 2024). Transformer based models such as Sentence-BERT have proven highly effective in generating semantic vector representations of sentences and documents, which can then be processed using vector-based search engines like FAISS. This integration allows search systems to operate in a more contextual, efficient, and scalable manner, making it possible to handle large document collections while maintaining accuracy and relevance. The implementation of semantic search in academic repositories not only enhances the relevance of search results but also contributes significantly to the overall quality of repository services. By interpreting the meaning behind a query, the system can provide results that better reflect user intent, even when the terminology differs from that used in the documents. This directly addresses the shortcomings of conventional keyword based approaches, which often fail to retrieve relevant materials due to lexical mismatches (Gao et al., 2021; Kulkarni et al., 2023).

Moreover, semantic search improves accessibility, allowing users from diverse academic backgrounds to obtain the information they need without having to adapt their search style to the linguistic structure of the documents. Ultimately, greater relevance supports research efficiency, accelerates literature discovery, and maximizes the use of available academic resources. By generating outputs aligned with user intent, academic repositories evolve beyond simple storage platforms into intelligent systems that actively support knowledge discovery. In the long run, such advancements contribute to building a more adaptive, inclusive, and responsive digital ecosystem for academic communities. Building on this background, the present study focuses on implementing semantic search within an academic repository by employing Sentence-BERT as the text embedding model and FAISS as the vector-based search engine. The system is evaluated using a range of performance metrics including Recall@K, NDCG, MRR, MAP, and latency to determine how effectively the proposed approach addresses the limitations inherent in traditional keyword-based search methods.

METHOD

Research Stages

This study was carried out through a series of systematic stages designed to ensure that the development and evaluation of the semantic search system within an academic repository could be conducted in a structured and measurable way. Each stage was interconnected, beginning with data collection and preparation, followed by the design of the semantic search model, and concluding with system performance evaluation. By following this approach, the study aims to provide a comprehensive assessment of the effectiveness of integrating Sentence-BERT and FAISS in the context of academic repositories.

The initial stage of this study began with problem analysis and a comprehensive literature review, aimed at identifying gaps in administrative document retrieval systems within academic environments while also building both theoretical and technical understanding of relevant technologies. This review included an exploration of embedding based text representation methods such as Sentence-BERT, vector similarity search techniques using FAISS, and existing digital repository systems.

In the next stage, the system architecture was designed, encompassing the formulation of the workflow, the organization of document storage structures, and the integration of key components such as the embedding model, FAISS indexing mechanism, and user interface. The design process emphasized efficiency, scalability, and user accessibility. Following the architectural design, the study advanced to dataset collection and preparation, which involved compiling representative administrative campus documents including decrees, circular letters, and other official records. These documents were digitized and subjected to text preprocessing to ensure consistency before being transformed into vector representations.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

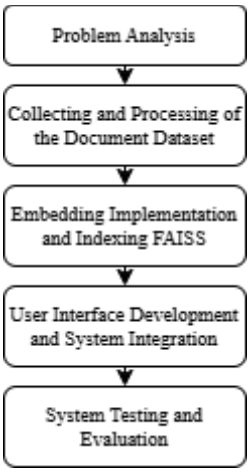

Figure 1. Research Methodology

During the document embedding phase, the entire corpus was processed using a deep learning–based embedding model to generate high dimensional vector representations enriched with semantic information. These representations allow the system to perform meaning-based retrieval rather than relying solely on literal keyword matching. Following this step, a search index was constructed using FAISS to enable efficient similarity search across the document collection.

The next stage involved developing a user interface that allows users to submit queries in natural language. The system processes these inputs by converting them into embeddings, matching them against the FAISS index, and returning a ranked list of semantically relevant documents. Finally, system testing and evaluation were carried out to measure the performance of the semantic search system and assess its effectiveness in meeting the intended objectives. The dataset and evaluation queries are written in Indonesian, reflecting the actual language used in the institutional repository. In order to ensure objective and reproducible evaluation, the relevance of documents to each query was defined using a binary relevance judgment scheme. A document was considered relevant if its textual content substantially addressed the intent expressed in the query, either directly through explicit terminology or indirectly through contextual semantic alignment. The relevance labeling process was conducted manually by domain experts familiar with the institutional repository content. Each query in the evaluation set was examined against the full document collection, and documents were marked as relevant if they satisfied at least one of the following criteria:

1. The document explicitly contains information that answers or addresses the query topic.
2. The document discusses the same administrative event, policy, or announcement referenced in the query, even if phrased differently, or
3. The document shares strong semantic alignment with the query intent despite lexical variation.

Documents that only contained overlapping keywords without contextual alignment were not considered relevant. This approach ensures that evaluation reflects semantic correctness rather than simple lexical matching.

Dataset

The dataset used in this study consists of 100 academic documents collected from the institutional repository. Although the dataset size is relatively moderate, it was intentionally selected to emphasize document diversity rather than sheer quantity. The documents represent various types of administrative content, including official letters, academic announcements, circulars, and guidelines, ensuring semantic variation across topics and writing styles. This design allows the study to focus on evaluating the effectiveness of semantic similarity modeling under realistic institutional conditions. Future research may extend the evaluation to larger scale repositories to further assess scalability performance. All data were organized in a MySQL database, structured to include key attributes such as id_documents, category, title, file type, file name, file path, and textual content (text_content). Among these fields, text_content serves as the primary focus of the study, as it contains the actual document content to be processed for semantic search purposes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1 Document Dataset

ID	Title	Type	FileName	Size (Byte)
1	Bakti Sosial Dies 8	pdf	Bakti Sosial Dies 8 - Lamp.pdf	512716
2	Surat Edaran Infaq Ramadhan 1446 H	pdf	Edaran Infaq Ramadhan 1446 H.pdf	333286
3	Pemberitahuan PMB pada Kegiatan Wisuda 06 November 2024	pdf	Pemberitahuan PMB pada Kegiatan Wisuda 06 November 2024.pdf	224925
4	Pengumuman Jadwal Mengajar di Bulan Ramadhan	pdf	Pengumuman Jadwal Mengajar di Bulan Ramadhan.pdf	292156
5	Pengumuman Libur dan Cuti Bersama Kenaikan Yesus Kristus	pdf	Pengumuman Libur dan Cuti Bersama Kenaikan Yesus Kristus 2024.pdf	378874
6	Pengumuman Libur Hari Raya	pdf	Pengumuman libur hari raya.pdf	470711
7	Pengumuman Libur Pilkada	pdf	Pengumuman Libur Pilkada.pdf	279837
8	Pengumuman Pindah Ruangan Sementara	pdf	Pengumuman Pindah Ruangan Sementara.pdf	212356
9	Pengumuman Teknis Pelaksanaan UAS Genap TA Untuk Dosen Pengampu Mata Kuliah. 20232024	pdf	Pengumuman Teknis Pelaksanaan UAS Genap TA Untuk Dosen Pengampu Mata Kuliah. 20232024.pdf	346122
10	Surat Penyampaian Laporan BKD Semester Genap TA. 20232024	pdf	Surat Penyampaian Laporan BKD Semester Genap TA. 20232024.pdf	723405
…	…	…	…	…
100	Pemberitahuan terkait Layanan BKD, Kenaikan Jabatan Non-PNS dan Kebutuhan Pemadanan Data	pdf	Surat Pemberitahaun Layanan BKD, Kenaikan Jabatan,Pemadanan Data.pdf	2452802

The dataset presents several challenges, particularly the presence of HTML tags within the text content, which result from the use of a WYSIWYG editor during document input. This condition necessitates a preprocessing stage to clean the text before it can be used as input for the embedding model.

Text Embedding Model (S-BERT)

Sentence-BERT (S-BERT) is a modification of the original BERT model designed to produce fixed length vector representations, or sentence embeddings, for sentences and documents (Gardazi et al., 2025). The key distinction from standard BERT lies in the addition of a pooling layer, which aggregates token level representations into a single vector. This enhancement enables the computation of semantic similarity between sentences or documents using distance or vector similarity measures, making S-BERT particularly effective for tasks that require meaning-based comparisons. In general, given a sentence S consisting of tokens $\{w_1, w_2, \dots, w_n\}$, BERT maps each token into a hidden representation:

$$h_i = BERT(w_i), i = 1,2, \dots, n \qquad (1)$$

To obtain a sentence representation, S-BERT applies a pooling function f over all token representations:

$$u = f(h_1, h_2, \dots, h_n) \qquad (2)$$

The pooling function can employ a simple mean operation:

$$u = \frac{1}{n}\sum_{i=1}^{n} h_i \qquad (3)$$

In this study, each academic document stored in the database was processed through several stages to generate its semantic vector representation. The first stage involved text preprocessing, which included removing HTML tags, eliminating special characters, and normalizing the text to ensure compatibility with the model's input requirements. Once cleaned, the text was fed into the Sentence-BERT model, where each word was tokenized and processed through the transformer encoder mechanism. This step produced contextual representations for each token based on bidirectional relationships within the sentence. Finally, a pooling layer was applied to aggregate the token representations into a single vector, effectively capturing the overall meaning of the sentence or document.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Vector Indexing and Search (FAISS)

FAISS (Facebook AI Similarity Search) is a library developed to perform similarity searches on high dimensional data (Douze et al., 2025). In general, the goal of FAISS is to identify the query vector x_q that most closely resembles a set of document vectors $\{x_1, x_2, \ldots, x_n\}$. This search is carried out by calculating similarity scores or distances between vectors. In this study, Euclidean distance (L2) is used as the primary measurement as follows :

$$d(x_q, x_i) = \|x_q - x_i\|_2 \qquad (4)$$

In this study, each academic document was first transformed into a vector representation using Sentence-BERT, and then indexed using FAISS. The indexing process involved storing all document embeddings within an IndexFlatL2 structure. This index type was selected because it is well suited to the dataset size and prioritizes search accuracy, which is essential for reliable semantic retrieval. IndexFlatL2 performs exact nearest neighbor search using Euclidean distance without approximation. This choice was made deliberately to prioritize retrieval accuracy and evaluation consistency over scalability optimization. Given the relatively moderate dataset size (100 documents), the computational cost of exact search remains negligible, and approximate nearest neighbor (ANN) methods such as IndexIVFFlat or HNSW do not provide significant practical benefits in terms of latency reduction at this scale. When a user submits a search query, the system converts the query into a vector embedding using the same Sentence-BERT model to ensure consistency. The resulting query embedding is then compared against the FAISS index to identify the top-k nearest documents based on vector similarity. The search results are returned as a ranked list of documents along with their corresponding similarity scores, and these results are subsequently displayed to the user through the repository system interface.

System Architecture

The development of the semantic search system for the academic repository was designed using a modular approach, ensuring that each component has a clearly defined role and can be developed or improved independently. The system architecture integrates a database layer, a text embedding model, a vector search engine, and an API-based interface service.

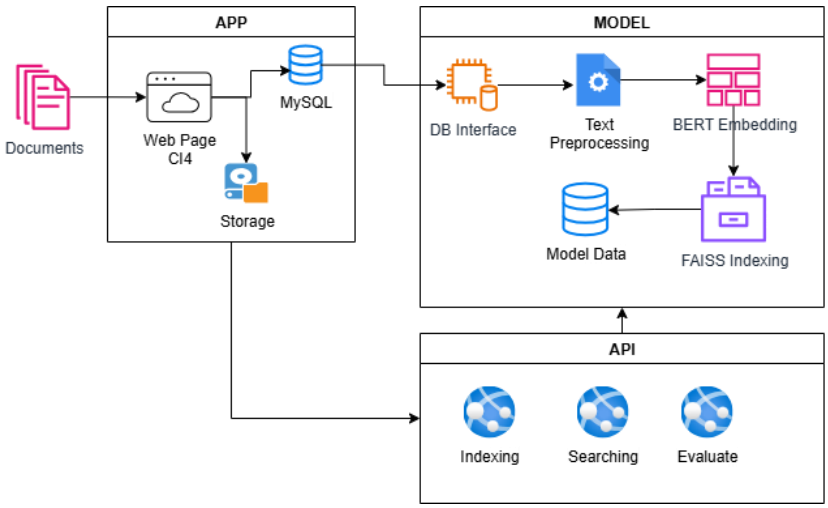

Fig. 2 System Architecture

The developed system architecture consists of three primary layers: the application layer, the model layer, and the API layer. At the application layer, academic documents are stored in a MySQL database and managed through an interface that enables both users and administrators to add, update, and access documents. The model layer serves as the core processing unit of the system. It begins with a database interface that connects the database to the processing modules, followed by the generation of text embeddings using Sentence-BERT to produce semantic vector representations of the documents. These vectors are then stored as structured model data and indexed using FAISS to enable similarity based semantic search. Meanwhile, the API layer provides structured services for indexing, searching, and evaluation, which can be accessed by the frontend application developed using

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

CodeIgniter 4. Through this integrated architecture, the system efficiently transforms raw document data into relevant search results, ensuring a smooth and effective semantic retrieval process for users.

RESULT

The implementation of the semantic search–based academic repository system was carried out by integrating backend, frontend, and database components in accordance with the previously defined architectural design. The backend was developed using FastAPI, which provides structured services for indexing, searching, and evaluation. These services are directly connected to the Sentence-BERT model for generating text embeddings and to FAISS as the vector search engine. Document data are stored in a MySQL database, which functions as the primary data repository. Meanwhile, the frontend interface was developed using CodeIgniter 4 to ensure convenient access for both administrators and end users.

On the administrative side, the system provides features for managing academic documents, including adding new records, updating existing content, and removing documents that are no longer relevant. Textual content is entered through a form editor that supports HTML formatting. Before proceeding to the embedding stage, the content undergoes automatic preprocessing to clean and normalize the text. Additionally, administrators can manually trigger the indexing process to ensure that newly added documents are immediately available within the semantic search system.

Web App

The web interface was developed to provide users with structured and user-friendly access for managing academic documents and utilizing the semantic search functionality. The frontend implementation was carried out using the CodeIgniter 4 (CI4) framework, selected for its lightweight architecture, support for the Model–View–Controller (MVC) pattern, and strong integration capabilities with the MySQL database. Meanwhile, interaction with the semantic search model is handled through a FastAPI-based backend built in Python. The primary objective of the web interface development was to ensure that all system functionalities could be operated intuitively by both system operators and general users. Operators are granted access to document management pages, model configuration pages, and evaluation data pages to maintain system sustainability and performance. In contrast, general users primarily interact with the search dashboard, which presents document retrieval results using a semantic search approach.

Fig. 3 Search Page

Overall, the web interface consists of four main pages:

1. Document Management Page (Document List) – Displays the list of documents available in the repository and provides functionalities to add, edit, deactivate, or delete documents.
2. Model Page – Presents information about the embedding model and the FAISS index currently in use, along with controls for re-indexing and system evaluation.
3. Evaluation Data Page – Manages evaluation datasets, including test sets, queries, and document relevance annotations used to measure search performance.
4. Search Page (Dashboard) – Provides a search input field for general users to retrieve documents using a semantic search approach.

Model Implementation

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

After completing the web interface implementation, the next stage involved evaluating the developed semantic search model. The primary objective of this evaluation was to assess how effectively the system delivers relevant search results in response to user needs, while also measuring its operational efficiency under real world conditions. The evaluation was conducted using three evaluation sets defined in the eval_sets table. Each set consists of multiple test queries designed to resemble natural user questions. For every query, the corresponding relevant documents (ground truth) were predefined in the eval_relevants table, enabling objective and reproducible metric computation.

The main evaluation metrics used in this study are as follows:

1. Recall@K. Measures the proportion of relevant documents successfully retrieved within the top K search results (Patel et al., 2022).
2. NDCG@K (Normalized Discounted Cumulative Gain). Evaluates the ranking quality of search results by considering the position of relevant documents (Yang et al., 2025).
3. MRR (Mean Reciprocal Rank). Calculates the average reciprocal rank of the first relevant document, reflecting how quickly a relevant result appears (Xing, 2024).
4. MAP (Mean Average Precision). Computes the average precision across all relevant positions for each query, then averages the result over all queries (Wang, Huang, et al., 2024).
5. No-Hit Rate – The percentage of queries that fail to return any relevant documents.
6. Latency (p50, p95, p99). System response time at the 50th, 95th, and 99th percentiles, serving as indicators of service speed and stability.
7. Embedding Time – The total time required to generate embeddings for all queries within a single evaluation set.

Table 2
Test Set

Set Name	Query	Target
tes 1	Ada pengumuman tentang juni 2025 pemberitahuan?	4 Juni 2025, Surat Pemberitahuan Perkuliahan pada hari Kamis, 05 Juni 2025, dilaksanakan secara online -LMS
tes 1	Ada pengumuman tentang juni 2025 pemberitahuan?	Pengalihan Kegiatan Proses Belajar Mengajar (PBM) Di Hari Rabu s.d. Kamis /12 s.d. 13 Juni 2024
tes 1	Ada pengumuman tentang juni 2025 pemberitahuan?	Penggunaaan LMS dalam Perkuliahan Semester Gasal TA. 2024/2025 di Fakultas Teknik dan Komputer – UnHar Medan
…	…	…
tes 3	Ada pengumuman tentang pemberitahuan terkait layanan?	4 Juni 2025, Surat Pemberitahuan Perkuliahan pada hari Kamis, 05 Juni 2025, dilaksanakan secara online -LMS
tes 3	Ada pengumuman tentang pemberitahuan terkait layanan?	Pemberitahuan PMB pada Kegiatan Wisuda 06 November 2024
tes 3	Ada pengumuman tentang pemberitahuan terkait layanan?	Pemberitahuan terkait Layanan BKD, Kenaikan Jabatan Non-PNS dan Kebutuhan Pemadanan Data

The first evaluation was conducted using 10 test queries with the parameter set to top_k = 10. The results indicate that the model was able to retrieve a substantial portion of relevant documents, although the distribution of relevance across the ranking positions still has room for improvement. A Recall@K value of 0.55 suggests that approximately 55% of the relevant documents were successfully retrieved within the top ten search results. Meanwhile, an NDCG@10 score of 0.409 indicates that the ranking quality remains suboptimal, as relevant documents do not consistently appear at higher positions in the result list.

Table 3. Test Results

Metrics	Test Set 1	Test Set 2	Test Set 3
Query Count	10	10	10
Top-K	10	10	10
Recall@K	0.55	0.667	0.6
NDCG@K	0.409	0.44	0.376

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15940
e-ISSN : 2541-2019
p-ISSN : 2541-044X

MRR	0.575	0.381	0.418
MAP	0.45	0.381	0.341
No-Hit Rate	0.1	0	0
Latency P50 (ms)	0.32	0.31	0.32
Latency P95 (ms)	0.44	0.43	0.43
Latency P99 (ms)	0.48	0.47	0.46
Total Embed Time (ms)	125.93	118.38	131.92

Interestingly, the MRR value (0.575) is relatively high, indicating that the first relevant document frequently appears at the top of the search results. This is further supported by a MAP score of 0.450, suggesting that the system's average precision is at a reasonably good level. However, the system still recorded a No-Hit Rate of 10%, meaning that one out of ten queries failed to retrieve any relevant documents. From an efficiency standpoint, this evaluation demonstrated extremely fast search latency, with p50, p95, and p99 values below 1 ms. The total embedding time for all queries was 125.93 ms, which remains computationally lightweight for a set of ten queries.

The second evaluation (Test Set 2) demonstrated improved retrieval coverage, with Recall@K increasing to 0.667 and NDCG@K rising to 0.440. This suggests better distribution of relevant documents within higher ranking positions. Despite the higher recall, the MRR decreased to 0.381, indicating that the first relevant document did not consistently appear at the very top. Importantly, the No-Hit Rate was reduced to 0, reflecting complete retrieval success across all queries in this set. In third evaluation, the system exhibited relatively balanced performance, with Recall@K at 0.600 and NDCG@K at 0.376. While recall remained stable compared to the previous sets, the slightly lower ranking metrics suggest room for further refinement in result ordering. The MRR (0.418) and MAP (0.341) indicate moderate ranking precision, and no retrieval failures were recorded.

Across all evaluation sets, latency values (p50, p95, and p99) consistently remained below 1 ms, demonstrating efficient response time under the tested conditions. Embedding time per set ranged from approximately 118 ms to 132 ms, indicating that the semantic embedding process remains computationally lightweight for small to medium-sized query batches.

DISCUSSIONS

The evaluation results indicate that the proposed semantic search system achieves consistent contextual retrieval performance across all test sets. An average Recall@K of 0.61 demonstrates that the embedding-based approach effectively retrieves a substantial portion of relevant documents within the top ten results. This confirms that Sentence-BERT successfully captures semantic relationships beyond exact keyword overlap, a capability particularly valuable in academic repositories where terminology varies across documents. However, the moderate NDCG values reveal that retrieval coverage does not always translate into optimal ranking quality. Although relevant documents are identified, they are not consistently positioned at the highest ranks. This distinction highlights a structural characteristic of dense embedding retrieval: semantic similarity ensures inclusion but does not inherently optimize ranking order. The observed variation between Recall, MRR, and MAP further reflects this trade-off. In some cases, higher recall coincides with lower MRR, indicating broader retrieval coverage at the expense of top-rank concentration. Such behavior aligns with known patterns in vector-based retrieval systems, where similarity scores may cluster semantically related documents without sharply separating the most relevant item. From a methodological perspective, the dataset scale (100 documents) allows evaluation of semantic effectiveness under realistic institutional conditions while minimizing computational noise. However, the limited corpus size and query volume restrict statistical generalization. The findings therefore validate semantic modeling capability rather than large-scale scalability performance. Future studies should assess robustness under larger and more heterogeneous repositories to examine ranking stability and approximation trade-offs.

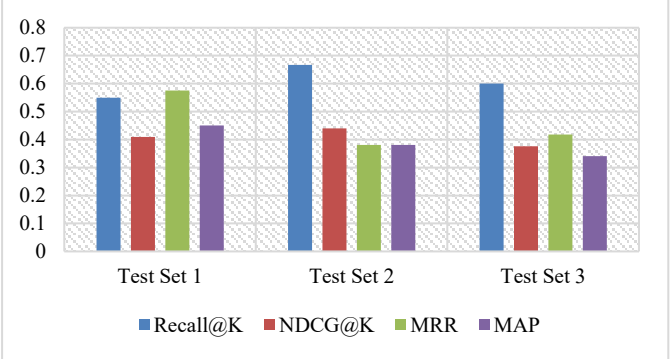

Fig. 4 Retrieval quality comparison

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

System efficiency results strengthen the practical relevance of the approach. Consistently low latency confirms that exact FAISS indexing (IndexFlatL2) does not impose performance constraints at the evaluated scale. This suggests that embedding-based retrieval can be integrated into institutional repositories without requiring advanced infrastructure. The results demonstrate that semantic search significantly improves contextual retrieval in academic repositories, particularly in handling synonym variation and natural-language queries. While ranking precision could be enhanced through hybrid or re-ranking strategies, the current implementation establishes a solid foundation for scalable, context-aware repository systems.

CONCLUSION

This study successfully implemented an academic repository system equipped with semantic search capabilities through the integration of Sentence-BERT and FAISS. The developed system addresses the limitations of conventional keyword-based search by providing more relevant and context aware retrieval results. Through a structured pipeline consisting of text preprocessing, embedding generation, vector indexing, and backend–frontend integration, the system can be effectively operated by both administrators and end users. Evaluation results across three test sets demonstrate that the system delivers consistent search performance, with an average Recall@K of 0.61, a MAP score of 0.39, and a No-Hit Rate of 0.033, indicating that most queries successfully retrieved at least one relevant document. Although the NDCG value showed a slight decline in the third test set, the system maintained strong capability in retrieving relevant documents within a relatively short response time. These findings confirm that the integration of Sentence-BERT for semantic representation and FAISS for similarity-based retrieval constitutes an effective solution for improving information search quality within academic repositories. Overall, this research contributes to the development of a more intelligent, adaptive, and user oriented academic repository system. The implementation of semantic search has been shown to enhance the retrieval experience by emphasizing contextual understanding rather than mere keyword matching. Future work may extend this research by exploring more advanced embedding models, incorporating approximate nearest neighbor techniques to improve scalability for large scale datasets, and conducting broader evaluations on more diverse academic document collections.

ACKNOWLEDGMENT

The authors gratefully acknowledge to Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) for the financial support provided through the 2025 Beginner Lecturer Research Grant under Contract No. 025/SK-M/VI/R.UnHar/2025.

REFERENCES

Acharya, S., Sornalakshmi, K., Paul, B., & Singh, A. (2022). Question Answering System using NLP and BERT. *3rd International Conference on Smart Electronics and Communication, ICOSEC 2022 - Proceedings*, 925–929. https://doi.org/10.1109/ICOSEC54921.2022.9952050

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. *Applied Sciences (Switzerland)*, *13*(6), 3911. https://doi.org/10.3390/app13063911

Diana, D., & Ekasari, M. H. (2021). Manajemen Tata Kelola Sistem Informasi Dokumentasi Surat Bagian Administrasi Umum Perguruan Tinggi. *Jurnal Ilmiah Komputasi*, *20*(1), 109–116. https://doi.org/10.32409/jikstik.20.1.2702

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). the Faiss Library. *IEEE Transactions on Big Data*. https://doi.org/10.1109/TBDATA.2025.3618474

Gao, L., Dai, Z., Chen, T., Fan, Z., Van Durme, B., & Callan, J. (2021). Complement Lexical Retrieval Model with Semantic Residual Embeddings. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12656 LNCS*, 146–160. https://doi.org/10.1007/978-3-030-72113-8_10

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artificial Intelligence Review*, *58*(6), 1–49. https://doi.org/10.1007/s10462-025-11162-5

Ghali, M.-K., Farrag, A., Won, D., & Jin, Y. (2025). Enhancing knowledge retrieval with in-context learning and semantic search through generative AI. *Knowledge-Based Systems*, *311*, 113047.

Heriani, A. P. S., Wahyudi, I., & Marsehan, A. (2025). Aplikasi Mobile untuk Meningkatkan Efisiensi Administrasi Kampus Universitas PGRI Silampari. *Sudo Jurnal Teknik Informatika*, *4*(2), 64–74. https://doi.org/10.56211/sudo.v4i2.854

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Kadang, M., & Nasaruddin, N. (2025). Desain dan Implementasi Sistem Repositori Dokumen Akademik Universitas DIPA Makassar. *E-Jurnal JUSITI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, *14*(1), 13–25. https://doi.org/10.36774/jusiti.v14i1.1712

Karri, N., & Jangam, S. K. (2024). Semantic Search with AI Vector Search. *International Journal of AI, BigData, Computational and Management Studies*, *5*(2), 141–150. https://doi.org/10.63282/3050-9416.ijaibdcms-v5i2p114

Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., & Alshahrani, S. M. (2022). Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, *8*, e967. https://doi.org/10.7717/peerj-cs.967

Krisnawati, L. D., Mahastama, A. W., Haw, S. C., Ng, K. W., & Naveen, P. (2024). Indonesian-English Textual Similarity Detection Using Universal Sentence Encoder (USE) and Facebook AI Similarity Search (FAISS). *CommIT Journal*, *18*(2), 183–195. https://doi.org/10.21512/commit.v18i2.11274

Kulkarni, H., MacAvaney, S., Goharian, N., & Frieder, O. (2023). Lexically-Accelerated Dense Retrieval. *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 152–162. https://doi.org/10.1145/3539618.3591715

Naqvi, S. M. R., Ghufran, M., Varnier, C., Nicod, J. M., Javed, K., & Zerhouni, N. (2024). Unlocking maintenance insights in industrial text through semantic search. *Computers in Industry*, *157–158*, 104083. https://doi.org/10.1016/j.compind.2024.104083

Patel, Y., Tolias, G., & Matas, J. (2022). Recall@k Surrogate Loss with Large Batches and Similarity Mixup. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 7492–7501. https://doi.org/10.1109/CVPR52688.2022.00735

Safira, F. (2021). Kebijakan Open Access Repositori Institusi di Perpustakaan Perguruan Tinggi: Kajian Best Practice Studi Literature. *Pustakaloka*, *13*(1), 116–136. https://doi.org/10.21154/pustakaloka.v13i1.2457

Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H., & Tovar-Arriaga, S. (2022). Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences*, *12*(2), 270. https://doi.org/10.3390/brainsci12020270

Tupan, T., & Rahayu, R. N. (2022). Narrative review: faktor-faktor yang berpengaruh terhadap pertumbuhan repositori akses terbuka (open access repositories) di Indonesia. *Al-Kuttab : Jurnal Kajian Perpustakaan, Informasi Dan Kearsipan*, *4*(1), 18–28. http://103.189.235.125/index.php/Kuttab/article/view/4992

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys*, *56*(7), 1–33. https://doi.org/10.1145/3648471

Wang, J., Zeng, J., & Sheng, J. (2024). Enhancing and Accelerating Image-Text Retrieval with Knowledge Graphs and FAISS. *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 1–6.

Xing, L. (2024). Secure Official Document Management and intelligent Information Retrieval System based on recommendation algorithm. *International Journal of Intelligent Networks*, *5*, 110–119. https://doi.org/10.1016/j.ijin.2024.02.003

Xiong, H., Bian, J., Li, Y., Li, X., Du, M., Wang, S., Yin, D., & Helal, S. (2024). When Search Engine Services Meet Large Language Models: Visions and Challenges. *IEEE Transactions on Services Computing*, *17*(6), 4558–4577. https://doi.org/10.1109/TSC.2024.3451185

Xu, S., Zhang, C., & Hong, D. (2022). BERT-based NLP techniques for classification and severity modeling in basic warranty data study. *Insurance: Mathematics and Economics*, *107*, 57–67. https://doi.org/10.1016/j.insmatheco.2022.07.013

Yang, W., Chen, J., Zhang, S., Wu, P., Sun, Y., Feng, Y., Chen, C., & Wang, C. (2025). Breaking the Top- K Barrier: Advancing Top- K Ranking Metrics Optimization in Recommender Systems . *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3542–3552. https://doi.org/10.1145/3711896.3736866

Zhu, P., Lang, Q., & Liu, X. (2023). Word Embedding of Dimensionality Reduction for Document Clustering. *Proceedings of the 35th Chinese Control and Decision Conference, CCDC 2023*, 4371–4376. https://doi.org/10.1109/CCDC58219.2023.10327354

Zoupanos, S., Kolovos, S., Kanavos, A., Papadimitriou, O., & Maragoudakis, M. (2022). Efficient comparison of sentence embeddings. *ACM International Conference Proceeding Series*, 1–6. https://doi.org/10.1145/3549737.3549752

*name of corresponding author